De la sémantique des contenus à la sémantique des structures

Laurent Romary

De la sémantique des contenus à la sémantique des structures

Laurent Romary

INTRODUCTION

La large diffusion du métalangage XML a eu pour conséquence heureuse d'offrir d'un seul coup un cadre syntaxique unifié pour la représentation de données qui étaient jusqu'alors définies et gérées par des communautés de spécialistes relativement différentes. De l'astronome au généticien, en passant par le spécialiste de littérature grecque classique, tous peuvent trouver en ligne des données exprimées sous la même forme et donc manipulables avec les mêmes outils. Cependant, ce rapprochement s'accompagne d'une évolution, voire d'une remise en cause, de certains concepts fondamentaux attachés aux anciennes pratiques de représentation de l'information. Il en est ainsi dans le domaine documentaire au sens large du terme, où les préoccupations classiques de nature bibliographique apparaissent de plus en plus complémentaires d'un besoin croissant d'archiver intégralement les contenus sous une forme électronique. Pourtant on perçoit très vite, quand on souhaite par exemple répertorier un corpus d'ouvrages informatisés, que les structures de catalogage classiques offrent un cadre un peu étriqué pour, par exemple, gérer tous les problèmes de versions ou de niveaux d'annotation multiples que pose le matériau électronique.

Par ailleurs, le besoin d'identifier de façon précise les informations les plus pertinentes au milieu de la masse de documents disponibles tels quels sur la Toile a motivé plusieurs initiatives pour définir des cadres de description des contenus de ces documents. On cherche ainsi à fournir des informations sur le contenu d'un document, ou métadonnées, sans qu'il soit nécessaire d'accéder effectivement au contenu lui-même. Ces initiatives – on pense typiquement au Dublin Core – cherchent à répondre à des besoins de simplicité et de diffusion large qui sont à l'opposé des objectifs de formats documentaires tels qu'UNIMARC.

Enfin, un courant particulier s'est développé autour de la notion de Web sémantique : il caresse l'espoir de définir un cadre unifié qui permette d'accéder aux documents de la Toile par le biais de descripteurs reflétant leur contenu informationnel. On a vu ainsi apparaître des syntaxes dédiées à ce genre de représentations – schémas RDF, OIL-DAML, Topic Maps –, mais aussi des initiatives de création de bases de concepts permettant de décrire des domaines d'activité particuliers.

Dans ce contexte encore instable, l'objectif de ce chapitre est d'essayer d'élargir le cadre d'analyse pour tenter, d'une part, d'aborder le problème des descripteurs de structures de documents, et, d'autre part, de dégager des pistes de convergence en montrant que l'on peut s'appuyer sur une base conceptuelle unifiée pour décrire la variété des métadonnées que l'on souhaiterait associer aux bases documentaires. Bien que prospectif, ce chapitre s'appuiera sur des expériences concrètes déjà mises en œuvre dans le domaine des terminologies multilingues et proposera des pistes pour des activités de standardisation indispensables au développement de véritables applications.

Quels moyens pour identifier des documents sur la Toile

Métadonnées spécialisées

Même en ne considérant que les documents produits dans le cadre de domaines spécialisés (littéraire, scientifique, économique, etc.), on constate que la quantité d'information disponible est telle qu'il est souvent devenu impossible d'accéder à l'ensemble des fonds correspondants par le seul biais de moteurs généralistes tels que Google ou Altavista, qui reposent tous sur une simple recherche par mots clés. Il est très vite apparu que la seule possibilité d'offrir un accès fiable et peu bruité était de développer des métadonnées spécifiques, définies par les spécialistes du domaine considéré, et permettant d'identifier ces documents.

L'une des initiatives pionnières en la matière est bien évidemment la TEI (*Text Encoding Initiative* [1] [3]) qui, dès le tout début des années 1990, impose que tout document conforme aux directives qu'elle publie possède un en-tête contenant un ensemble structuré de descripteurs de son contenu [1]. Comme l'ensemble des directives de la TEI repose sur la norme SGML, cet en-tête peut s'exprimer sous la forme d'une structure hiérarchique reposant sur quatre grandes parties :
– une section dédiée à la description du fichier électronique (<fileDesc>, ou *file description*), qui contient toutes les informations relatives au document électronique proprement dit (celui qui a été balisé) et permet à un documentaliste de cataloguer ce document. Cette section contient aussi les informations relatives aux sources éventuelles du document (par exemple, la publication d'origine d'un texte qui aurait été numérisé) ;

– une section permettant de décrire les choix de transcription et / ou d'annotation adoptés dans le document par rapport à sa source (<encodingDesc>, *encoding description*). Ces informations portent autant sur le contenu (par exemple, la correction des erreurs dans la source ou la normalisation de l'orthographe pour des sources anciennes) que sur le balisage. Nous reviendrons juste après sur ce point ;
– une section qui regroupe les descriptions liées au contenu informationnel du document (<profileDesc>, *text profile*), et qui peut aussi bien contenir des mots clés, une description du sujet du texte ou, dans le cas par exemple de transcription de données audio, la liste des personnes qui interviennent. On peut noter ici qu'il est possible d'indiquer les langues utilisées dans le document (<langUsage>), information qui, comme on le verra, est à mi-chemin entre le contenu et la structure ;
– une section (<revisionDesc>, *revision history*) dédiée à la gestion des versions du document électronique en donnant la possibilité d'enregistrer l'histoire des révisions qu'il a subies.

L'annexe I (p. 227) montre un exemple d'en-tête TEI destiné à documenter un corpus de transcriptions de dialogues oraux.

La richesse des métadonnées contenues dans l'en-tête TEI a conduit au développement de nombreuses archives textuelles [2] exploitant les différents champs disponibles pour offrir des recherches complexes ou des navigations dans des bases de textes. Elle n'a cependant pas conduit à l'organisation de véritables réseaux internationaux d'archives de textes, car il manquait une infrastructure de recueil et de centralisation des métadonnées qu'un utilisateur aurait pu interroger par le biais d'un portail unique. Une telle infrastructure aurait nécessité le développement de mécanismes propres à la TEI, à un moment où le déploiement de SGML était loin d'atteindre celui que l'on connaît actuellement pour XML. Ce vide a en quelque sorte été comblé par des initiatives plus récentes profitant de la large diffusion de descripteurs génériques de métadonnées procurées par des propositions à portée plus large.

C'est ainsi qu'a vu le jour le projet OLAC (*Open Language Archive Community*) [3]. Ce projet est principalement issu de la communauté des linguistes, notamment de terrain, ayant à gérer de nombreux documents issus de recueils de données sur site et souhaitant par ailleurs connaître avec précision les corpus déjà disponibles pour telle ou telle langue. Dans un souci aigu de pragmatisme, la communauté OLAC a décidé de s'appuyer au maximum sur des propositions génériques de description de métadonnées et d'infrastructure d'accès à celles-ci. OLAC repose donc sur deux piliers fondamentaux :
– le Dublin Core [4] comme cadre de description des métadonnées associé à tout document répertorié dans le réseau de sites affiliés à OLAC. Plus particulièrement, les 15 descripteurs peuvent être associés à des facettes

affinant leur sémantique et / ou leur portée. Par exemple, OLAC introduit le champ <Subject.language>, variante du champ <Subject> du Dublin Core, pour indiquer la langue décrite par le document [5] ;
– l'OAI (*Open Archives Initiative*) [6], comme protocole de moissonnage des métadonnées. Le principe de base utilisé par l'OAI est que chaque site affilié à une initiative de métadonnées procure un accès à un fichier XML – disponible en statique, ou généré dynamiquement à la demande via un script CGI, par exemple – qui est régulièrement consulté par un portail centralisé.

Parmi les champs offerts par OLAC, on peut noter la possibilité de fournir des indications techniques relatives au format des ressources identifiées par les descripteurs (Format.markup). Les informations peuvent ainsi porter sur le type de la ressource (enregistrements d'échanges spontanés, textes, lexiques, etc.), sur le support des fichiers primaires audio ou vidéo, ou encore sur le type de codage ou de transcription utilisé. Il est même possible, quand l'information est pertinente (et surtout disponible), de donner une référence à un outil permettant de manipuler la ressource. L'annexe II (p. 229) montre un extrait simplifié (une seule entrée de texte) du fichier défini à l'ATILF pour la représentation de la base Frantext, en conformité avec les directives OLAC.

Des contenus aux structures

Le projet OLAC, même s'il permet de réunir sous une même bannière une multiplicité de ressources (électroniques ou non, d'ailleurs) dont la gestion est répartie sur le réseau, ne fait que déporter le problème quant à l'utilisation effective de celles-ci. De fait, OLAC n'est pas destiné à fournir un accès réel aux ressources, tout au plus une indication que tel ou tel document existe quelque part, accompagnée des principales caractéristiques de celui-ci. Si l'on veut effectivement manipuler la ressource linguistique qui se trouve derrière une description particulière, par exemple pour la visualiser, ou exploiter les annotations qui lui sont associées, il faudrait pouvoir connaître avec précision l'organisation du document, c'est-à-dire, dans le cas où XML serait la syntaxe de référence, la DTD ou le schéma sur lequel celui-ci repose. Dans ce cadre, OLAC va le plus souvent pointer sur une référence générale indiquant, par exemple, que le document est conforme aux directives de la TEI.

En fait, nous rencontrons là un problème bien plus général, celui de la multiplicité des structures pouvant exister au sein d'un même groupe de documents – ici, les ressources linguistiques. Bien plus, il arrive que, pour des objets de natures extrêmement similaires, des formats très différents soient utilisés. Comment, alors, identifier ces formats avec précision ? Comment garantir qu'un outil logiciel saura ou non traiter un format

particulier ? C'est en quelque sorte l'objet même de ce chapitre, et nous allons ici donner quelques éléments de réponses parmi celles qui ont été proposées jusqu'à présent.

De façon générale, connaître la structure d'un document correspond à partager une certaine connaissance relative à celle-ci entre un émetteur (ou producteur) et un récepteur (ou consommateur) de document. Si cette connaissance est partagée à priori, on parle alors d'« échange en aveugle ». Si, au contraire, elle a lieu au moment de la transaction, on parle d'« échange négocié ». Ce dernier type d'échange ne garantit pas, par définition, que l'information pourra être comprise à tout coup et nécessite, comme on le verra dans la partie suivante, des moyens conceptuels et techniques plus élaborés.

L'échange en aveugle peut s'opérer suivant deux modes. Le mode implicite repose sur l'hypothèse qu'émetteur et récepteur ont la même connaissance du format partagé, sans souci de vérification au moment de la transmission de l'information. Ce mode, qui est celui de la majorité des échanges de documents HTML sur le Web, requiert un fort taux de tolérance au niveau du destinataire qui n'a aucun moyen de garantir la conformité de la donnée qu'il reçoit au format annoncé. Le mode explicite est celui adopté par les différentes initiatives visant à créer des répertoires de DTD ou de schémas XML. Il suppose l'existence d'un espace centralisé où émetteur et récepteur vont trouver une même description de référence des structures qu'ils vont respectivement produire et consommer. Ainsi, Microsoft a ouvert un espace où toute personne ou projet peut déposer un schéma XML et y faire référence avec la garantie que ce schéma sera toujours accessible.

Le mode explicite des échanges en aveugle présente l'avantage immédiat qu'il devient possible de véritablement intégrer aux métadonnées d'un document une information précise quant à l'organisation structurelle de celui-ci. Cela permet par exemple à un logiciel d'effectuer un test de conformité de façon automatique. Il présente cependant le défaut majeur, déjà observé dans le cadre de l'initiative de Microsoft, de ne pas résoudre le problème de la profusion des formats pour un même type de document. En effet, toute variante, aussi petite soit-elle, conduit à la création d'une nouvelle entrée dans le référentiel. Le diagnostic de conformité est donc du type tout ou rien, alors même que l'on souhaiterait ignorer par exemple le balisage superflu, ou surtout accepter des documents correspondant à un sous-ensemble du format de référence reconnu par le destinataire.

Ce référentiel trop macroscopique privilégie donc le développement de schémas surgénérateurs contenant un maximum d'éléments, et ne répondant donc pas au besoin qu'ont des applications particulières de contrôler finement leurs données.

Vers une description plus fine des structures – retour sur la TEI

L'une des difficultés que l'on rencontre dès que l'on cherche à décrire plus précisément la structure des documents auxquels on souhaite donner accès en ligne est de déterminer si les descripteurs doivent s'adresser à un lecteur humain ou à un processus automatique. Bien souvent, ces deux types de destinataires sont d'ailleurs vus comme incompatibles : une documentation simplifiée mais explicite va satisfaire le lecteur ayant besoin, par exemple, de connaître le niveau d'annotation d'une ressource linguistique, tandis qu'un outil informatique aura besoin d'une description exhaustive, mais condensée, du document pour pouvoir effectuer son travail. Pour illustrer cette contradiction, nous pouvons nous tourner de nouveau vers la TEI, là encore pionnière en la matière.

L'une des caractéristiques les plus importantes de la TEI est que, malgré l'étendue des possibilités de codage qu'elle autorise, elle s'organise autour d'une DTD modulaire qui permet de ne sélectionner qu'un certain sous-ensemble des éléments disponibles. Ainsi, sur la base de quelques structures communes (le jeu de balise « noyau », dont fait partie l'en-tête), il est possible de sélectionner un type de texte particulier (les « bases » ; par exemple les balises pour le codage de la prose, de la poésie, du théâtre, des dictionnaires, etc.) ainsi que des modules additionnels de balisage (noms et dates, pointeurs, etc.).

Le modèle sous-jacent, dit de la « pizza de Chicago », est en fait assez grossier, car il n'est pas possible de choisir véritablement les éléments nécessaires à une application donnée. La configuration la plus simple contient ainsi par défaut une petite centaine d'éléments.

Du point de vue de la documentation des formats, cette modularité, même relativement simple, rend quasi impossible l'utilisation de répertoires de DTD ou de schémas, telle que nous l'avons décrite précédemment. On souhaiterait pouvoir reconnaître les propriétés communes à tous ces formats pour, au moins, savoir que l'on peut appliquer les mêmes feuilles de style, ou encore générer une base de données bibliographiques à partir des en-têtes compris dans l'ensemble des textes d'une archive.

Paradoxalement, la TEI n'offre pas réellement de mécanisme plus élaboré de documentation. L'en-tête d'un document donné permet bien, dans sa section <encodingDesc>, d'indiquer la liste des éléments effectivement utilisés dans le corps du document, par le biais de l'élément <tagsDecl> (*tagging declaration*, ou déclaration de balisage), mais cette description ne répond qu'à moitié aux besoins d'interopérabilité que nous avons exprimés jusqu'à présent.

Afin de mieux appréhender ce que permettent les déclarations de balisage dans l'en-tête TEI, considérons un exemple. Le document ci-dessous pourrait représenter un extrait de la déclaration de balisage d'un dictionnaire codé conformément aux directives de la TEI :

```
<tagsDecl>
    <tagUsage gi="div" occurs="26">Utilisé pour marquer les séparations alphabétiques du dictionnaire.</tagUsage>
    <tagUsage gi="entry" occurs="14526"/>
    <tagUsage gi="orth" occurs="22638"/>
    <tagUsage gi="sense" occurs="8304"/>
    …
</tagsDecl>
```

Cet exemple illustre clairement le fait qu'il n'est nulle part indiqué qu'il s'agit de la description d'un dictionnaire. De fait, l'ensemble des occurrences d'éléments est réuni dans une structure à plat qui mélange les éléments issus du noyau de la TEI et ceux propres au module de codage de dictionnaires (par exemple <entry>). On ne peut donc pas reconstituer la recette de la pizza à laquelle correspond le document à partir de cette description. Par ailleurs, comme la déclaration de balisage d'un document ne porte que sur le corps textuel de celui-ci [7], les balises utilisées dans l'en-tête lui-même ne sont pas répertoriées. Or, il est souvent intéressant de savoir avec quel degré de précision une ressource a été décrite (identification précise des locuteurs pour un recueil de données linguistiques sur le terrain, par exemple).

Au bilan, on observe que la TEI a privilégié un mode de description des structures orienté vers l'utilisateur humain qui saura en quelque sorte compléter l'information partiellement décrite. L'hypothèse est ici faite que les mécanismes propres à XML (les inclusions, en l'occurrence) sont suffisants pour un traitement automatique, bien qu'il n'y ait aucune cohérence entre ces deux possibilités de description.

En prenant un peu plus de distance, on observe que les mécanismes proposés par la TEI sont intimement liés au fait que celle-ci repose intégralement sur XML pour la description de ses documents et que, de ce fait, il existe un répertoire d'éléments parfaitement identifiés *au sein de la communauté TEI*. Ainsi, quand bien même les informations portées par l'en-tête seraient plus précises, elles ne rendraient la structure du document accessible qu'à un processeur connaissant à priori ce répertoire d'éléments.

Aucune garantie d'interopérabilité n'est de toute façon offerte avec d'autres types de documents. Il serait donc intéressant d'étendre la notion de descripteur de structures pour parvenir à dépasser le cadre spécifique d'un métalangage comme XML. C'est ce que nous allons tenter progressivement dans les parties suivantes.

MODÉLISER ET DOCUMENTER DES STRUCTURES

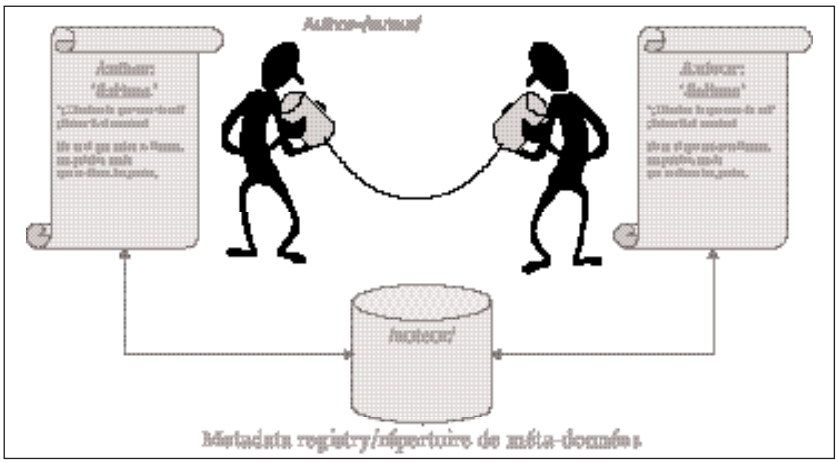
Perspective générale

Le survol des initiatives dans le domaine des métadonnées montre bien qu'il reste encore du chemin à parcourir avant de systématiser l'intégration d'informations permettant de reconstituer exactement l'organisation (par exemple, le balisage XML) d'un document donné. Cependant, avant de vouloir généraliser trop rapidement, peut-être est-il nécessaire de revenir sur la notion même de métadonnée, notamment sur un plan sémantique, et d'envisager de quelle façon il est possible de l'étendre aux informations structurelles.

Le principe d'une métadonnée est de fournir une information structurée quant au contenu d'un document destiné à être transmis d'une source A à un destinataire B (celui-ci n'étant pas nécessairement spécifié à priori, par exemple dans le cas du Web). Pour ce faire, il est nécessaire que le *vocabulaire* utilisé pour décrire ces métadonnées soit connu de la même façon par A et par B pour qu'effectivement un document annoncé comme ayant une certaine propriété soit reconnu comme tel par le récepteur potentiel (qui explore le Web à la recherche d'une certaine information). On retrouve ici l'argumentaire déjà développé pour la notion de modèle de document.

Bien que posé justement en terme de vocabulaire, le partage de métadonnées est un peu plus complexe. Il faut garantir en effet qu'un descripteur donné, par exemple <Author> sera compris de la même façon par A et par B. Ceci n'est possible que si le descripteur en question est en quelque sorte certifié par une autorité tierce, qui notamment en donnera une définition précise garantissant la portée exacte du concept ainsi représenté (l'auteur est-il l'auteur du contenu informationnel, celui qui a numérisé le document, etc. ?). Si l'on veut aller encore plus loin, on constate que la notion même de vocabulaire est restrictive, car rien ne garantit que toutes les communautés, linguistiques en particulier, souhaiteront utiliser les mêmes termes pour désigner les mêmes concepts. L'autorité doit alors devenir un véritable répertoire de métadonnées, permettant à chacun de mettre en relation ses propres descripteurs avec un concept central. Ce principe est illustré en figure 1 ci-contre, où les vocabulaires utilisés de part et d'autre du canal de communication (<Author> et <Auteur>) sont tous deux appariés au concept central ici étiqueté /auteur/ (le nom importe alors peu, il s'agit surtout d'avoir un identifiant unique pour le répertoire). Ainsi, au-delà de la notion de vocabulaire partagé, une métadonnée ne peut être comprise que si elle correspond au partage d'une véritable ontologie de concepts, que ceux-ci soient génériques (cas du Dublin Core), ou propres à un domaine d'application particulier (par exemple, OLAC [8]).

FIGURE 1. PARTAGE CONCEPTUEL DE MÉTADONNÉES PAR LE BIAIS D'UN RÉPERTOIRE CENTRALISÉ

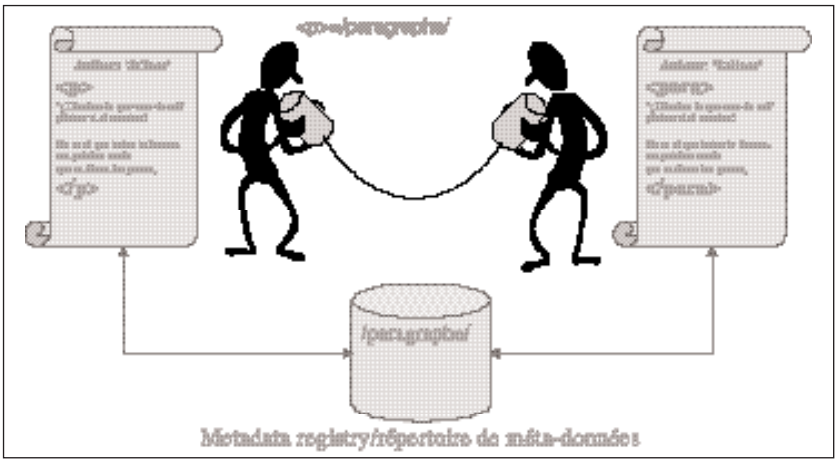


Cependant, cette vision implique que, d'un point de vue technique, l'émetteur A communique, en même temps que les informations relatives à son document, l'appariement de son propre vocabulaire avec les catégories du registre de référence, soit en transmettant, comme une sorte d'en-tête à ses métadonnées, la liste de ses appariements, de façon à donner au récepteur une clé d'interprétation de ses propres données, soit en neutralisant ses données (en les exprimant dans un format pivot indépendant d'un vocabulaire particulier et donc reposant uniquement sur les catégories abstraites du répertoire central). Bien que la seconde solution semble la plus facile à mettre en œuvre, on doit garder la première à l'esprit car elle correspond à toute situation où l'on accède à des données statiques propriétaires qui n'auraient justement pas été neutralisées.

Supposons maintenant qu'il ne s'agisse pas seulement d'identifier un document au regard d'un certain nombre de descripteurs mais, de façon plus fine, de comprendre l'organisation interne des données contenues dans ce document. La figure 2 (en page suivante) présente ainsi une situation où l'émetteur du document utilise un balisage XML particulier pour lequel le niveau du paragraphe est indiqué par l'élément `<p>`. De son côté, le récepteur souhaite intégrer ce document dans une base où les paragraphes sont indiqués à l'aide de l'élément `<para>`. En l'absence de tout autre mécanisme, une telle situation nécessiterait à l'émetteur et au récepteur de définir ensemble la correspondance entre ces deux éléments (et probablement quelques autres…) pour aboutir à un filtrage éventuel d'un format vers l'autre. Nous sommes donc dans une situation d'*échange négocié* qui alourdit toute transmission et surtout doit être réitéré à chaque nouvelle situation d'échange, ou pour tout nouveau type de document.

Si l'on suppose maintenant qu'il existe un registre de descripteurs de structures textuelles, c'est-à-dire un lieu où des notions telles que /paragraphe/ sont définies de façon précise, au sein d'une communauté, ou mieux au niveau d'une organisation internationale telle que l'ISO, la situation est quelque peu différente. La notion de paragraphe ayant été définie préalablement à la transmission, l'émetteur peut communiquer en même temps que son document un appariement entre éléments XML qu'il utilise et les catégories correspondantes au sein du registre. Un tel registre, pour peu qu'il contienne une base suffisamment large de descripteurs, décrit une véritable ontologie de structures documentaires contribuant à l'élaboration de la sémantique d'un document particulier [4].

Figure 2. Échange d'informations structurelles par le biais d'un répertoire de métadonnées



Là encore, une possibilité – que nous rencontrerons avec le langage GMT – est de neutraliser la structure du document en passant par un format pivot plus abstrait. Reste que l'on perçoit que le passage de la métadonnée simple à l'échange explicite d'informations relatives à la structure des documents semble être envisageable.

Le paysage n'est pourtant pas si rose qu'il paraît :
– nous n'avons pas encore les moyens de décrire de tels registres. C'est l'objet de la partie suivante ;
– nous n'avons pas encore de garantie que l'intuition de faisabilité de l'échange d'information sur les structures est valide. Nous nous appuierons sur l'expérience acquise dans le cadre de la plate-forme TMF (*Terminological Markup Framework*) pour avancer dans ce sens.

Métadescription des métadonnées

Comme nous l'avons vu, le partage d'une métadonnée correspond à l'existence d'un système conceptuel connu à la fois de l'émetteur et du récepteur de cette métadonnée. Ce système conceptuel, suivant le domaine d'origine de celui qui va l'aborder, sera vu alors de deux façons relativement différentes :
– il pourra être considéré comme la donnée d'une véritable ontologie de concepts, tels qu'on peut en rencontrer dans le domaine de la représentation des connaissances. On envisagera alors des langages élaborés pour la représenter, langages qui, dans le monde SGML/XML, prendront les noms de Topic Maps, ou OIL-DAML. Une telle approche permet d'utiliser des outils puissants, tels que les logiques de description, pour faire des inférences sur les types d'objets, mais peut s'avérer lourde à mettre en œuvre ;
– il pourra être simplement vu comme un ensemble d'unités techniques qu'il faut créer, maintenir et rendre accessible de façon cohérente pour que le système de communication fonctionne. C'est typiquement l'approche adoptée par la norme ISO/IEC-11179 [5], et que nous allons détailler maintenant.

L'ISO/IEC-11179 est une norme internationale en six parties qui propose un cadre de description d'éléments de données (*data element*), c'est-à-dire de champs pouvant être utilisés pour exprimer des métadonnées. Il s'agit donc d'un format de « méta-métadonnées ». L'objectif affiché est, d'une part, de permettre une description consistante des éléments de données d'une application à une autre (en fournissant, par exemple, des directives précises concentrant la formulation des définitions), et, d'autre part, de fournir des méthodes d'enregistrement et de mise à jour de ces éléments de données.

Dans sa troisième partie, l'ISO/IEC-11179 introduit en particulier un certain nombre d'attributs pouvant être utilisés pour la description de l'élément de donnée considéré. Les principaux attributs sont rappelés dans le tableau 1, en page suivante (le nom anglais des attributs est donné entre parenthèses).

Ainsi le Dublin Core a-t-il décidé d'adopter le cadre de la norme ISO-11179 pour décrire son ensemble d'éléments de métadonnées [2]. L'élément <Creator> y est par exemple décrit tel que dans le tableau 2, en page suivante.

Sur cette base, un projet plus spécifique tel qu'OLAC, mais gardant le Dublin Core comme base, va pouvoir raffiner certains champs. Dans le cas de <Creator> par exemple, OLAC va compléter les commentaires, ajouter des attributs et des exemples d'utilisation (tableau 3, p. 215).

Tableau 1. Liste des principaux attributs proposés par la norme ISO/IEC 11179 pour la description de métadonnées

Nom (*Name*)	Étiquette assignée à l'élément de donnée
Identifiant (*Identifier*)	L'unique identifiant assigné à l'élément de donnée
Version (*Version*)	La version de l'élément de donnée
Autorité d'enregistrement (*Registration Authority*)	L'entité autorisée à répertorier l'élément de donnée
Langue (*Language*)	Langue dans laquelle est décrit l'élément de donnée
Définition (*Definition*)	Un énoncé qui représente clairement le concept correspondant à l'élément de donnée
Obligation (*Obligation*)	Indique si l'élément doit être obligatoirement présent ou non (c'est-à-dire doit posséder une valeur déterminée)
Type de donnée (*Datatype*)	Indique le type de la donnée qui peut être utilisée comme valeur de l'élément de donnée
Occurrence maximale (*Maximum Occurrence*)	Indique le nombre maximum de fois que l'élément de donnée peut être répété
Commentaire (*Comment*)	Toute remarque additionnelle concernant l'utilisation de l'élément de donnée

Tableau 2. Description de l'élément <Creator> du Dublin Core sur la base des attributs de la norme ISO/IEC 11179

Nom	Creator
Identifiant	Creator
Version	1.1
Autorité d'enregistrement	Dublin Core Metadata Initiative
Langue	en
Définition	An entity primarily responsible for making the content of the resource
Obligation	Optional
Type de donnée	Character String
Occurrence maximale	Unlimited
Commentaire	Examples of a Creator include a person, an organisation, or a service. Typically, the name of a Creator should be used to indicate the entity

Tableau 3. Descriptions spécifiques associées à l'élément de donnée <Creator> par OLAC

Commentaires	A Creator may be a person, an organization, or a service. Creator is closely related to Contributor. In determining whether an entity is a Creator (as opposed to a Contributor), use the same criteria that are followed for deciding that an entity should be listed in the «author» slot of a bibliographic reference as a primary source of the intellectual content. Entities that do not merit that level of recognition should be treated as Contributors. Recommended best practice is to identify a Creator by means of a name and to give the name in a form that is ready for sorting within an index. For the names of persons, this means that the name should be given in inverted order with the surname first. For the names of organizations, this means that any initial article should be omitted. When a resource has more than one creator, use a separate Creator element for each one.
Attributs	The refine attribute is optionally used to specify the role (such as author, editor, translator, and so on) played by the named entity in the creation of the resource.
Exemples	A personal author: <creator>Bloomfield, Leonard</creator> An institutional author: <creator>Linguistic Society of America</creator> An editor: <creator refine="editor">Sapir, Edward</creator>

Clairement, la perspective adoptée est encore très fortement inspirée par les descriptions de vocabulaires spécialisés. Il reste à formaliser un peu plus ces descriptions, et l'on peut espérer, à terme, une convergence entre une telle norme et des approches plus liées à la notion d'ontologie.

Application – la plate-forme TMF

Nous allons maintenant montrer, dans le cas d'une application particulière dédiée à la représentation des terminologies multilingues, comment il est possible de mettre en œuvre des mécanismes intégrant les éléments de réflexion précédents avec, d'une part, une modélisation de structures à l'aide de descripteurs génériques et, d'autre part, une représentation de ces descripteurs dans une implémentation en RDF d'un sous-ensemble de l'ISO/IEC-11179.

Principes généraux

Nous reprenons ici les principaux éléments formant la base de la plate-forme TMF (*Terminological Markup Framework*) [9], qui devrait devenir le futur standard ISO-16642 de description de données terminologiques informatisées. L'objectif sous-jacent à la proposition de cette norme était double. Tout d'abord, il s'agissait de décrire et de comparer les formats d'échange terminologiques existants tels que MARTIF [6] ou Geneter [8], tant en terme de couverture descriptive que pour identifier les conditions d'interopérabilité entre ces formats. Une telle démarche a conduit le groupe qui a travaillé sur ce texte à définir des principes plus généraux permettant d'analyser les bases de données existantes (par exemple, la base Eurodicautom [10] de l'Union européenne) et de les mettre en correspondance avec un format d'échanges de données particulier.

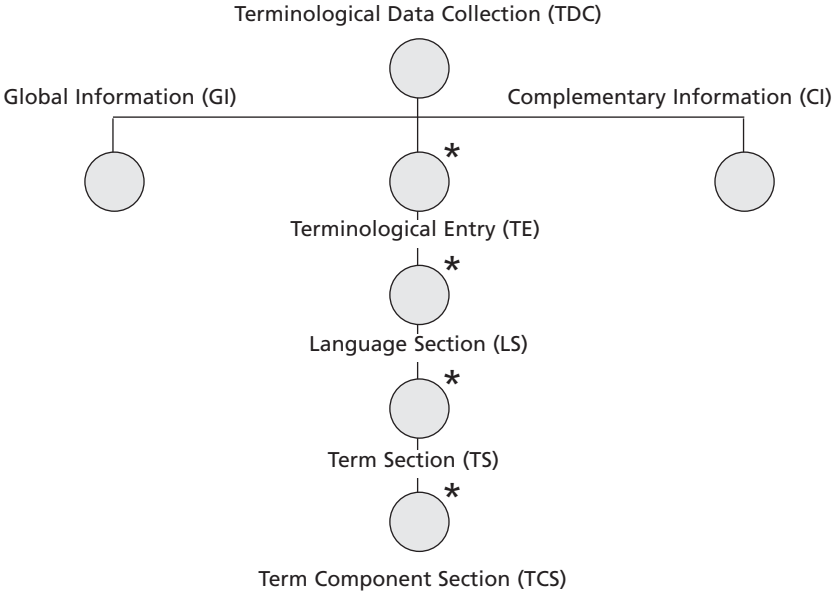
Par ailleurs, il s'agissait de répondre à un besoin pressant de l'industrie de disposer de plus de flexibilité dans la définition de ses propres formats, tout en gardant un maximum de compatibilité avec les normes existantes. L'idée était ainsi de pouvoir définir un format terminologique simple ne reposant que sur cinq unités d'information exprimées à l'aide d'éléments et d'attributs XML (<id>, <générique>, <terme>, <catégorie>, et <langue>), conformément à l'extrait suivant [11], tout en se gardant la possibilité de comparer (ou de compléter) ces données avec celles issues d'Eurodicautom :

```
<?xml version="1.0" encoding="iso-8859-1"?>
<base>
    <notion>
        <id>1</id>
        <générique target="#notion[id='2']">Insecte</générique>
        <expression langue="fr">
            <description>
                <terme>Abeille</terme>
                <catégorie>Nom</catégorie>
            </description>
        </expression>
        <expression langue="en">
            <description>
                <terme>Bee</terme>
                <catégorie>Nom</catégorie>
            </description>
        </expression>
    </notion>
</base>
```

La plate-forme TMF permet de décrire un format quelconque pour l'échange de terminologies multilingues en utilisant, par exemple, XML comme syntaxe d'implémentation. Ainsi, TMF ne décrit aucun format particulier, mais fait office de langage de spécification, sur la base des notions élémentaires suivantes :

– un métamodèle : une organisation générale des données partagée par tous les formats [12] pouvant être engendrés par la plate-forme TMF, et qui décompose la structure d'une base de données terminologiques en composants élémentaires tels que montré sur la figure 3. Ce modèle est conforme à la vision traditionnelle d'une entrée terminologique comme une composante d'un réseau conceptuel, vision remontant aux travaux déjà anciens d'Eugen Wüster [9] et largement adoptée dans la communauté de la terminologie ;
– des unités élémentaires d'information (ou catégories de données), qui sont extraites d'un répertoire de catégories de données (DCR – *Data Category Registry*) en fonction des besoins de l'application courante. La spécification de l'ensemble des catégories de données d'un format particulier peut impliquer la description de catégories propriétaires, ce qui rend le format non interopérable avec les autres ;
– des méthodes d'implémentation permettant de décrire l'ensemble des caractéristiques du format considéré (ou TML) en instanciant le métamodèle et les catégories de données à l'aide de vocabulaires idoines (par exemple pour décrire des éléments ou attributs XML, ou encore des champs dans une base de données relationnelle).

Figure 3. Le métamodèle d'une base de données terminologique



TDC, Collection de données terminologiques ; GI, Informations globales ; CI, Informations complémentaires ; TE, Entrée terminologique ; LS, Section de la langue ; TS, Section du terme ; TCS, Section du composant du terme.

Étude d'un exemple

Pour illustrer les principes de TMF, nous pouvons chercher à identifier les éléments précédents dans la décomposition d'une entrée terminologique typique exprimée dans le format particulier TBX, une variante du format MARTIF (ISO 12200 [6]) adoptée par l'industrie de la localisation [13] :

```
<termEntry id='ID67'>
    <descrip type='subjectField'>manufacturing</descrip>
    <descrip type='definition'>A value between 0 and 1 used in ...
    </decrip>
    </langSet lang='en'>
        <tig>
            <term>alpha smoothing factor</term>
            <termNote type='termType'>fullForm</termNote>
        <tig>
    </langSet>
    </langSet lang='hu'>
        <tig>
            <term>Alfa simitisi tenyezo</term>
        <tig>
    </langSet>
<termEntry>
```

Dans cet exemple, on peut distinguer deux types d'objets représentés dans la structure XML. D'une part, un certain nombre d'éléments (marqués en gras) organisent une sorte de squelette structurel reflétant le métamodèle terminologique : <termEntry> (correspondant au niveau *Terminological Entry*), <langSet> (*Language Section*), <tig> (*Term Section*). Ces éléments n'apportent aucune information particulière ; ils sont juste destinés à organiser l'entrée terminologique de façon hiérarchique. Les objets XML restants, qu'ils soient exprimés à l'aide d'éléments (<term>), d'éléments dits « typés » (<descrip>, <termNote>) ou d'attributs (<id>, <lang>), contribuent directement à la description de l'entrée terminologique et peuvent être mis en correspondance avec les catégories de données issues de la norme ISO-12620 [7], comme montré ci-dessous.

Tableau 4. Mise en correspondance des objets XML de l'exemple TBX-Martif avec les catégories de données de la norme ISO-12620

Objet Martif	Style	Identifiant de la catégorie (ISO 12620)	Nom de la catégorie (ISO 12620)
<term>	Element	ISO12620-A01	Term
<descrip type='subjectField'>	Typed element	ISO12620-A04	Subject field
<descrip type='definition'>	Typed element	ISO12620-A0501	Definition
<termNote type='termType'>	Typed element	ISO12620-A0201	Term type
'id'	Attribute	ISO12620-A1015	Entry identifier
'lang'	Attribute	ISO 12620A100701	Language Identifier

Une représentation possible de cette décomposition peut être obtenue en utilisant un format intermédiaire pivot. Dans le cas de TMF, ce format, appelé GMT (*Generic Mapping Tool*), repose sur deux éléments principaux [14] : <struct>, qui instancie le métamodèle terminologique et <feat>, qui permet d'attacher les descripteurs aux différents niveaux du métamodèle. On obtient ainsi :

```
<struct type="TE">
    <feat type="id">ID67</feat>
    <feat type="subjectField">manufacturing</feat>
    <feat type="definition">A value between 0 and 1 used in
    ...</feat>
    <struct type="LS">
        <feat type="lang">en</feat>
        <struct type="TS">
            <feat type="term">alpha smoothing factor</feat>
            <feat type="termType">fullForm</feat>
        </struct>
    </struct>
    <struct type="LS">
        <feat type="lang">hu</feat>
        <struct type="TS">
            <feat type="term">Alfa simitisi tenyezo</feat>
        </struct>
    </struct>
</struct>
```

Interopérabilité entre deux formats de représentation de terminologies

Étant données deux TML conformes aux spécifications de la plate-forme TMF, l'expression des conditions d'interopérabilité se réduit à la comparaison de leurs spécifications respectives en termes de catégories de données, de par le fait qu'elles partagent exactement le même métamodèle [15]. De fait, les méthodes de spécification de la plate-forme TMF permettent de faire un diagnostic précis de la quantité d'information qui sera préservée ou perdue lorsque l'on passe d'un TML à un autre, ce que l'on nomme la « bande passante d'interopérabilité ».

Pour illustrer ce concept, nous pouvons comparer l'exemple simplifié que nous avons présenté en p. 216, correspondant à un premier TML (TML_1), et l'extrait de la base exprimé en TBX (ci-contre), format compatible avec TMF (que nous nommerons TML_2). Nous ferons l'hypothèse que les formats TML_1 et TML_2 sont définis sur la base des spécifications minimales permettant de couvrir les deux exemples de ce chapitre, notamment en termes de catégories de données utilisées. Dans un premier temps, nous pouvons associer à chaque objet XML du format TML_1 les catégories de données correspondantes de la norme ISO-12620, comme nous l'avions fait

pour l'exemple TBX. Le résultat est présenté dans le tableau 5, qui est une mise en correspondance des objets XML de l'exemple de la p. 216 (TML_1) avec les catégories de données de la norme ISO 12620.

Tableau 5. Mise en correspondance des objets XML de l'exemple TML_1 (p. 216) avec les catégories de données de la norme ISO-12620

Objet Martif	Style	Identifiant de la catégorie (ISO 12620)	Nom de la catégorie (ISO 12620)
<terme>	Element	ISO12620-A01	Term
<générique>	Element	ISO12620-A070201	Broader concept generic
<catégorie>	Element	ISO12620-A020201	Part of speech
<id>	Element	ISO12620-A1015	Entry identifier
langue	Attribute	ISO 12620-A100701	Language Identifier

Nous pouvons alors aligner les deux spécifications de catégories de données des formats TML_1 et TML_2 (tableau 6) pour identifier la bande passante d'interopérabilité qui, ici, est limitée à trois unités d'information.

Tableau 6. Identification de la bande passante d'interopérabilité entre les formats TML_1 et TML_2

Catégorie de donnée du format TML_1	Interopérabilité	Catégorie de données du format TML_2
Term	Compatible	Term
	Perte de TML_2 vers TML_1	*Subject field*
	Perte de TML_2 vers TML_1	*Definition*
	Perte de TML_2 vers TML_1	*Term type*
Entry identifier	Compatible	Entry identifier
Language Identifier	Compatible	Language Identifier
Broader concept generic	*Perte de TML_1 vers TML_2*	
Part of speech	*Perte de TML_1 vers TML_2*	

Styles et vocabulaires

L'évaluation de la bande passante d'interopérabilité entre deux TML repose uniquement sur les catégories de données proprement dites, et se trouve donc entièrement indépendante de l'implémentation de celles-ci en tant qu'objets XML. Par exemple, la catégorie de données /Entry identifier/ est instanciée comme un attribut dans le format TML_2 (TBX) et comme un élément dans le

format TML_1, sans que cela remette en cause le transfert de l'information correspondante de l'un vers l'autre. C'est pour cette raison que la plate-forme TMF possède un mécanisme complémentaire, mais disjoint, permettant de décrire la façon dont un TML sera concrètement implémenté comme type de document XML, une fois que la spécification des catégories de données a été effectuée. Il s'agit en fait d'associer à chaque catégorie de données un style [16] qui sélectionne une réalisation possible en XML, d'une part, et un vocabulaire, correspondant aux chaînes de caractères nécessaires à l'expression de ce style, d'autre part. Par exemple, le format TML_1 implémente la catégorie de données /Broader concept generic/ de la façon suivante :

```
Style       Element
Vocabulary  "générique"
```

Cela permet d'utiliser un élément <générique> pour représenter l'unité d'information correspondante.

Représentation des catégories de données en RDF

Comme nous l'avons vu, la comparaison de deux TML n'est possible que s'il existe un registre central de catégories de données, où celles-ci sont représentées de façon consistante. Nous avons également vu comment la norme ISO/IEC-11179 pouvait servir de cadre à la formalisation d'un tel registre. Cependant, afin que le processus de spécification et de comparaison de TML puisse être automatisé, la plate-forme TMF s'appuie [17] sur une formalisation accrue de l'ISO/IEC-11179 qui exprime ses différents attributs à l'aide du format RDF (*cf.* [10]). Cette représentation présente l'avantage de permettre à un utilisateur qui souhaite spécifier un format conforme à TMF d'ajouter ses propres catégories de données à celles qu'il aura extraites d'un registre de référence.

La figure 4 (page suivante) montre le modèle RDF utilisé pour définir de façon unique une catégorie de données. On notera quelques modifications par rapport aux attributs de base de l'ISO/IEC-11179 pour relier une catégorie de données à un ou plusieurs niveaux du métamodèle (*Level*), spécifier si la catégorie de données correspond à une valeur simple ou à un champ (*DCType*), ou relier les catégories de données entre elles par une relation de parenté (*DCParent*). Le modèle RDF contient les attributs nécessaires à l'expression des styles et vocabulaires associés à chaque catégorie de données.

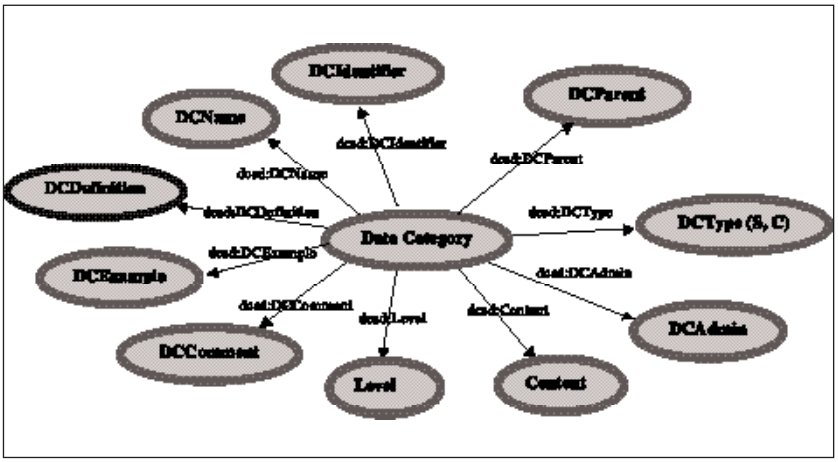
Les catégories de données vues comme métadonnées

Sans entrer plus avant dans les détails techniques, on peut anticiper que les mécanismes décrits précédemment ouvrent le chemin à l'automatisation de

nombreux traitements. Tout d'abord, une spécification complète (contraintes sur les catégories de données, styles et vocabulaires) permet de générer le schéma XML d'un TML ainsi décrit.

Cette même spécification permet de générer un filtre XSLT permettant de transformer des données d'un TML donné vers le format pivot GMT (et inversement), et, par transitivité, de générer les filtres de transformation entre deux TML (dans la limite de la bande passante d'interopérabilité).

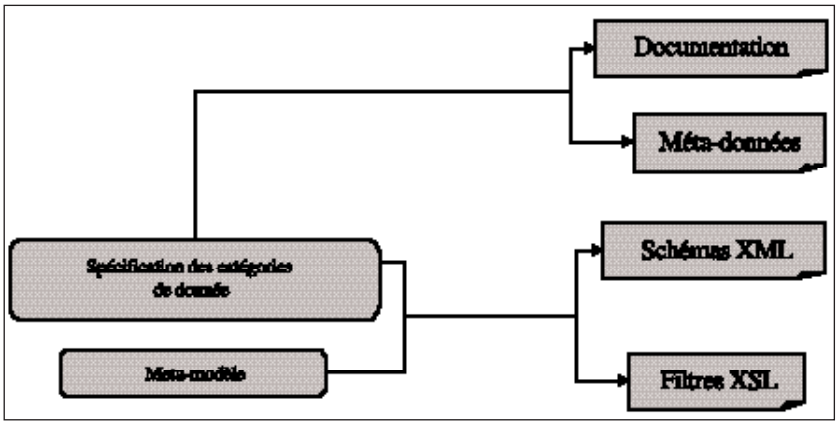
Figure 4 : Principaux attributs RDF utilisés pour la description des catégories de données



Comme on le voit dans la figure 5 ci-contre, on aboutit à une intégration plus étroite des fonctions de documentation et de manipulation d'un document. La même spécification utilisée dans le cadre de la plate-forme TMF permet en effet de contrôler, en amont, les formats des documents qui seront créés ou manipulés, et elle sert de base, en aval, à une documentation précise de ces formats qui, puisqu'elle est structurée, peut être intégrée dans un système de métadonnées.

Dès lors, tout transfert de données terminologiques peut être accompagné d'une référence au fichier de spécifications correspondant, donnant ainsi une information précise sur le contenu structurel de ces données. Sans même regarder les données, le récipiendaire peut alors effectuer un diagnostic sur le degré de compatibilité de celles-ci avec sa propre base et effectuer en conséquence les traitements nécessaires.

Figure 5 – Place de la spécification des catégories de données dans les différentes opérations de manipulation et de diffusion des documents



BILAN ET PERSPECTIVES DANS UN CONTEXTE DE NORMALISATION DES FORMATS DE DONNÉES

Nous avons essayé de présenter dans ce chapitre différents éclairages destinés à intégrer la représentation de la structure de documents dans le champ plus global des métadonnées. Ce faisant, nous avons pu voir que cela, d'une part, conduisait à une réflexion de fond sur l'organisation et le partage de métadonnées au travers de registres internationaux de référence, et permettait, d'autre part, d'aborder le problème relativement complexe de la spécification de formats de documents, notamment dans le cadre des possibilités offertes par le métalangage XML. Bien que limitée à un type de données très particulier, l'expérience acquise dans le cadre de la plate-forme TMF semble donner l'espoir qu'il soit possible, pour certaines classes de document (qui restent à identifier), de généraliser les mécanismes présentés ici.

Ceci est particulièrement important dans un contexte où il existe des forces antagonistes au sein de nombreuses communautés qui souhaitent à la fois normaliser de façon de plus en plus précise les données qu'elles échangent, et conserver une certaine souplesse dans la description même de ces formats. Il n'est donc plus question de normaliser au niveau international un schéma XML particulier, mais bien de fournir les moyens aux différents acteurs d'une communauté de comparer leurs usages et de maintenir éventuellement les formats de données qu'ils manipulaient précédemment.

Il reste probablement encore beaucoup de chemin à parcourir pour que toute application puisse être intégrée dans cette perspective. Il faut en effet identifier les métamodèles correspondants, ajouter éventuellement certains mécanismes que le domaine des terminologies ne nécessitait pas, mettre en œuvre des répertoires thématiques de catégories de données, et probablement définir des modes plus rigoureux de gestion de ces catégories de données sur la base de l'expérience acquise en représentation des connaissances. Il semble malgré tout que ce soient là des activités indispensables si l'on veut accroître les niveaux d'intéropérabilité entre applications dans les années à venir.

NOTES

1. À propos de la TEI, voir aussi le chapitre 3 de cet ouvrage.

2. On peut citer par exemples le Women Writers Project (http://www.wwp.brown.edu/), le projet Silfide (http://www.loria.fr/projets/Silfide) ou encore le Perseus Project (http://www.perseus.tufts.edu/), exemplaire par ailleurs pour sa navigation multilingue.

3. http://www.language-archives.org/

4. http://dublincore.org/documents/dces/

5. Langue qui peut, bien sûr, différer de celle utilisée dans le document lui-même. On peut très bien identifier une grammaire en anglais portant sur un créole parlé en Afrique, par exemple.

6. http://www.openarchives.org/

7. La documentation de la TEI rappelle que : « *A* <tagsDecl> *must contain exactly one occurrence of a* <tagUsage> *element for each distinct element marked within the outermost* <text> *element, associated with the* <teiHeader> *in which it appears.* »

8. On peut aussi mentionner les travaux réalisés par le projet européen ISLE pour la définition d'un ensemble de métadonnées pour les ressources linguistiques, qui va bien au-delà, en finesse et en précision, du projet OLAC. C'est en particulier la nécessité d'unifier les descripteurs d'OLAC et de ISLE qui conduit l'ISO, dans le cadre de son nouveau comité TC37/SC4, à mettre en place un répertoire conceptuel unique de métadonnées pour les ressources linguistiques.

9. *Cf.* http://www.loria.fr/projets/TMF/ pour d'autres documents, échantillons de données et fichiers de transformation XSLT.

10. *Cf.* http://eurodic.ip.lu/

11. Pour des raisons pédagogiques, la « base de données » est limitée à une seule entrée. Cette entrée fait cependant référence à une autre entrée qui intervient dans une relation spécifique-générique par le biais d'un lien XML ("#notion[id='2']". Voir http://www.w3.org/XML/Linking).

12. Ou TML, *Terminological Markup Language*.

13. *Localisation Industry Standard Association* (http://www.lisa.org).

14. Afin de rendre compte de toutes les possibilités descriptives de TMF, le format pivot GMT comprend d'autres éléments, notamment <brack>, qui permet de regrouper des descripteurs qui s'interdéfinissent, et <annot>, qui permet d'intégrer des annotations dans le contenu d'un descripteur. GMT contient aussi plusieurs attributs génériques pour la représentation des langues (xml:lang) et des liens (source et target).

15. La proposition de norme ISO/CD 16642 identifie plus précisément les conditions liées aux valeurs (types) des catégories de données, à l'ancrage de celles-ci sur les nœuds du métamodèle, etc. Ainsi, une définition exprimée au niveau de l'entrée terminologique (TE) dans un TML ne sera pas intégrable à un modèle qui n'accepte celle-ci qu'au niveau de la section langagière (LS).

16. TMF autorise cinq styles (*Attribute, Element, Typed Element, Valued Element, Typed Valued Element*), pour couvrir les différentes possibilités rencontrées dans les formats en terminologie.

17. Ce modèle sert de base à la partie 1 de la proposition de révision de la norme ISO-12620.

Références

[1] Association for Computers and the Humanities (ACH), Association for Computational Linguistics (ACL), and Association for Literary and Linguistic Computing (ALLC) ; C. M. Sperberg-McQueen, L. Burnard, eds. Guidelines for Electronic Text Encoding and Interchange (TEI P3). 2 vol. Chicago, Oxford, Text Encoding Initiative, 1994

[2] Dublin Core Metadata Element Set, Version 1.1: Reference Description. Dublin Core Metadata Initiative, 1999. http://dublincore.org/documents/dces/

[3] N. Ide, J. Veronis, eds. The Text Encoding Initiative: Background and Contexts. Special triple issue of Computers and the Humanities, 1995, vol. 29, n° 3

[4] C. Weltis, N. Ide. Using the right tools: enhancing retrieval from market-up documents. Journal of Computers and the Humanities, April 1999, vol. 33, n° 10, p. 59-84

[5] ISO/IEC 11179 – Specification and Standardization of Data Elements, Parts 1-6. Genève, Organisation internationale de normalisation
ISO/IEC 11179-1: Framework for the Specification and Standardization of Data Elements. 1999
ISO/IEC 11179-2: Classification for Data Elements. 2000

ISO/IEC 11179-3: Basic Attributes of Data Elements. 1994
ISO/IEC 11179-4: Rules and Guidelines for the formulation of Data Definitions. 1995
ISO/IEC 11179-5: Naming and Identification Principles for DEs. 1995
ISO/IEC 11179-6: Registration of Data Elements. 1997

[6] ISO 12200 – Applications informatiques en terminologie – Format de transfert de données terminologiques exploitables par la machine (MARTIF) – Transfert négocié. Genève, Organisation internationale de normalisation, 1999

[7] ISO 12620 – Aides informatiques en terminologie – Catégories de données. Genève, Organisation internationale de normalisation, 1999

[8] A. Le Meur. GENETER: a generic format for the distribution and reuse of heterogeneous multilingual data, Proc. LREC, Grenada, 1998

[9] H. Picht, K.-D. Schmitz, eds. Terminologie und Wissensordnung, Ausgewählte Schriften aus dem Gesamtwerk von Eugen Wüster. TermNet Publisher, 2001

[10] Resource Description Framework (RDF) Model and Syntax Specification, W3C Recommendation, 22 February 1999. http://www.w3.org/TR/REC-rdf-syntax/

ANNEXE I - EN-TÊTE TEI DÉCRIVANT LE CORPUS « SNCF » (GDR-PRC COMMUNICATION HOMME-MACHINE)

```
<teiHeader date.updated="27/06/2002" date.created="27/06/2002" sta
tus="new" creator="Jean-François MAURET" type="corpus">
  <fileDesc>
    <titleStmt>
      <title>corpus SNCF: version électronique en français</title>
      <respStmt>
        <resp>Encodeur</resp>
        <name>
          <persName>
            <foreName>Jean-François</foreName>
            <surname>Mauret</surname>
          </persName>
        </name>
      </respStmt>
    </titleStmt>
    <extent>263 textes de communications répartis en 3 phases
</extent>
    <extent>phase 1: 117 textes (caller/operator)</extent>
    <extent>phase 2: 85 textes (caller/machine)</extent>
    <extent>phase 3: 61 textes (caller/machine)</extent>
    <extent> 915 ko </extent>
    <publicationStmt >
      <distributor>
        <orgName>
          <orgDivn>
            <name>Equipe LED</name>
          </orgDivn>
          <name>Loria</name>
          <address>
            <addrline>Campus Scientifique</addrline>
            <addrline>BP 239</addrline>
            <addrline>54506 VANDŒUVRE-LES-NANCY</addrline>
          </address>
        </orgName>
      </distributor>
      <authority>
        <persName>
          <foreName>Laurent</foreName>
          <surname>Romary</surname>
        </persName>
      </authority>
    </publicationStmt>
    <sourceDesc>
      <recordingStmt>
        <recording type="audio">
```

```
        <p>Enregistrement de conversations téléphoniques entre des
étudiants et une opératrice ou une machine</p>
        </recording>
      </recordingStmt>
    </sourceDesc>
  </fileDesc>
  <encodingDesc>
    <editorialDecl>
      <normalization>
        <p>TEI P4</p>
      </normalization>
    </editorialDecl>
    <projectDesc >
      <p>Corpus recueilli dans le cadre du GDR –PRC Communication
Homme-Machine</p>
    </projectDesc>
  </encodingDesc>
  <profileDesc>
    <particDesc >
      <person>
        <p>student</p>
      </person>
      <person>
        <p>machine</p>
      </person>
      <person>
        <p>operator</p>
      </person>
      <person>
        <p>caller</p>
      </person>
    </particDesc>
  </profileDesc>
  <revisionDesc>
  …
  </revisionDesc>
</teiHeader>
```

Annexe II - Extrait d'un fichier de métadonnées conforme au format défini par OLAC

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE OLAC-Repository SYSTEM "oryx.dtd">
<OLAC-Repository>
  <identity>
    <repositoryIdentifier>ATILF:FRANTEXT</repositoryIdentifier>
    <repositoryName>ATILF Resources</repositoryName>
    <adminEmail>mailto:olac-admin@inalf.fr</adminEmail>
    <olac-archive type="institutional">
      <archiveURL>http://atilf.inalf.fr/frantext.htm</archiveURL>
      <institution>CNRS-ATILF</institution>
      <institutionURL>http://www.inalf.fr/atilf</institutionURL>
      <location>BP 30687, 54063 Nancy cedex, FRANCE</location>
      <synopsis>textual database</synopsis>
    </olac-archive>
  </identity>
  <records>
    <record spec="all">
      <header>
        <recordId>oai:ATILF:FRANTEXT:M277</recordId>
        <datestamp>2002</datestamp>
      </header>
      <metadata>
        <olac>
              <coverage>19th century</coverage>
              <creator refine="author">Hugo, Victor</creator>
              <date refine="created" code="1920"/>
              <description lang="fr"> Nouvelle édition publiée par Paul
              Berret avec le concours de l'Académie Française. Saisie inté-
              grale. </description>
              <identifier lang="fr">Hugo, V, La légende des siècles, tome 1
              </identifier>
              <language code="fr"/>
              <publisher>Paul Berret Ed, Paris</publisher>
              <relation refine="isPartOf">oai:ATILF:FRANTEXT</relation>
              <rights>free</rights>
              <subject.language>fr</subject.language>
              <title>La legende des siecles, tome 1 </title>
        </olac>
      </metadata>
    </record>
    ...
  </records>
</OLAC-Repository>
```